Solutions to Assignment #9

1. Suppose you want to test whether $1,000$ observations come from a normal$(3, 4)$ distribution. The observations are divided into the following ranges:

$$\begin{aligned} A_1 &= \{x \in \mathbb{R} \mid x \leqslant 0\} \\ A_i &= \{x \in \mathbb{R} \mid i - 2 < x \leqslant i - 1\} \quad \text{for } i = 2, 3, \ldots, 7, \text{ and} \\ A_8 &= \{x \in \mathbb{R} \mid x > 6\}. \end{aligned}$$

It is recorded that $60, 96, 140, 210, 172, 160, 88$, and 74 of the observations fall into $A_1, A_2, \ldots, A_8$, respectively. Do the data support the assertion that the observations come from a normal$(3, 4)$ distribution?

(a) Compute the probability, p_i, that a given observation falls into the A_i range.

Solution: Table shows that categories $A_1, A_2, \ldots, A_8$ with their corresponding probabilities, $p_1, p_2, \ldots, p_8$, which were computed using the NORMDIST function in MS Excel and whose format is

$$\text{NORMDIST}(\text{x}, \text{mean}, \text{sd}, \text{TRUE}),$$

where the TRUE gives the cdf for normal random variable with expected value mean and variance sd^2.

A_i (i)	p_i	Predicted Counts	Observed Counts
1	0.0668	66.8	60
2	0.0918	91.8	96
3	0.1499	149.9	140
4	0.1915	191.5	210
5	0.1915	191.5	172
6	0.1499	149.9	160
7	0.0918	91.8	88
8	0.0668	66.8	74

Table 1: Categories A_i

For example, for i=1 ,2,...,7,

$$p_i = \text{NORMDIST}(i - 1, 3, 2, \text{TRUE}) - \text{NORMDIST}(i - 2, 3, 2, \text{TRUE}).$$

□

(b) Compute the expected counts in each category under the null hypothesis, H_o, that $p_1, p_2, \ldots, p_8$ describe the true distribution of counts.

Solution: The expected counts are shown in the third column in Table 1. □

(c) Compute the Pearson Chi–Square statistic for the data given here.

Solution: The value of the Pearson Chi–Square statistic, Q, is given by

$$\widehat{Q} = \sum_{i=1}^{8} \frac{(X_i - E_i)^2}{E_i}$$

where X_i denotes the observed count in category i and E_i the corresponding expected count. Using the values in the third and fourth columns in Table 1 we obtain

$$\widehat{Q} \approx 6.92$$

□

(d) Would you reject H_o at the 5% level of significance?

Solution: The Pearson Chi–Square statistic, Q, has an approximate $\chi^2(7)$ distribution in this case. It then follows that the approximate p–value for this test is

$$p\text{–value} = \text{P}(Q \geqslant \widehat{Q}) \approx 0.4373,$$

which is much larger than 0.05. Thus, we cannot reject H_o at the 5% level of significance. □

2. A die was cast $n = 120$ independent times and the following data resulted If we

Spots Up	1	2	3	4	5	6
Frequency	b	20	20	20	20	$40 - b$

use a chi–square test, for what values of b would the hypothesis that the die is unbiased be rejected at the 0.025 significance level?

Solution: In this case, the null hypothesis is

$$H_o: \quad p_i = \frac{1}{6} \quad i = 1, 2, \ldots, 6.$$

Thus, the expected values are

$$E_i = 20 \quad i = 1, 2, \ldots, 6.$$

The value of the Pearson Chi–Square statistic, Q, is then

$$\widehat{Q} = \frac{(b - 20)^2}{20} + \frac{(40 - b - 20)^2}{20} = \frac{(b - 20)^2}{10}.$$

In this case Q has an approximate $\chi^2(5)$ asymptotic distribution. Thus, for the approximate p–value to be less than 0.025, we require that

$$\frac{(b - 20)^2}{10} \geqslant 12.833,$$

according to the Chi–Square Distribution table on page 673 in the text. Solving for b we then obtain that

$$b \geqslant 32 \quad \text{or} \quad b \leqslant 8.$$

$\square$

3. Consider the classical problem from Mendelian genetics of crossing two types of peas. The Mendelian theory states that the probabilities of the following classifications

   (a) round and yellow;
   (b) wrinkled and yellow;
   (c) round and green; and
   (d) wrinkled and green

are $\dfrac{9}{16}, \dfrac{3}{16}, \dfrac{3}{16},$ and $\dfrac{1}{16},$ respectively.

Suppose that from 160 independent observations the frequencies of the respective classifications are 86, 35, 26 and 13. Are these data consistent with the Mendelian theory? Justify your answer; that is, state the hypothesis you are testing and indicate the significance level that you using to make your decision.

   Solution: The third column in Table 2 shows the expected counts for this situation based on the null hypothesis

   $$\mathrm{H}_o: \quad p_a = \frac{9}{16}, \quad p_b = \frac{3}{16}, \quad p_c = \frac{3}{16}, \quad p_d = \frac{1}{16}.$$

Category	Probabilities	Expected Counts	Observed Counts
(a)	9/16	90	86
(b)	3/16	30	35
(c)	3/16	30	26
(d)	1/16	10	13

Table 2: Data for Problem 3

Thus, the value of the Pearson Chi–Square statistic, Q, for these data is

$$\widehat{Q} = 2.44.$$

In this case Q has an approximate $\chi^2(3)$ asymptotic distribution. Hence, the approximate p–value for this test is

$$p\text{–value} = \mathrm{P}(Q \geqslant \widehat{Q}) \approx 0.4862.$$

Thus, there is a large likelihood of seeing the observed values, or more extreme ones, under the assumption that the null hypothesis is true. Therefore, we cannot reject the null hypothesis on the basis of the data which were provided, and conclude that the data are consistent with the Mendelian theory. □

4. A certain genetic model suggests that the probabilities of a particular trinomial distribution are $p_1 = p^2$, $p_2 = 2p(1 - p)$ and $p_3 = (1 - p)^2$, respectively, where

$$0 < p < 1.$$

If X_1, X_2 and X_3 represent the respective frequencies in n independent trials, explain how we could check the adequacy of the genetic model.

Solution: The null hypothesis for this test is

$$\mathrm{H}_o: \quad p_1 = p^2, \quad p_2 = 2p(1 - p), \quad p_3 = (1 - p)^2.$$

The expected counts for each of the three categories are shown on Table 3.

It then follows that he value for the Pearson Chi–Square statistics for this test is

$$\widehat{Q} = \frac{(X_1 - np^2)^2}{np^2} + \frac{(X_2 - 2np(1 - p))^2}{2np(1 - p))^2} + \frac{(X_3 - n(1 - p)^2)^2}{n(1 - p)^2}.$$

Category (i)	p_i	Expected Counts	Observed Counts
1	p^2	np^2	X_1
2	$2p(1-p)$	$2np(1-p)$	X_2
3	$(1-p)^2$	$n(1-p)^2$	X_3

Table 3: Data for Problem 4

If the value of the parameter p is known, then, for large values of n, Q has an approximate $\chi^2(2)$ distribution. Using this information, we can then approximate the p–value

$$p\text{–value} = \mathrm{P}(Q \geqslant \widehat{Q}).$$

If p–value $< \alpha$, for some pre–set significance level, α, we then reject the null hypothesis, H$_o$. If the p–value is too big compared to a pre–assigned significance level, α, we would say that the observations are consistent with the proposed genetic model. $\square$

5. It is proposed to fit the Poisson distribution to the following data

x	0	1	2	3	$x \geqslant 4$
Frequency	20	40	16	18	6

(a) Compute the corresponding chi–square goodness of fit statistic

Suggestion: In estimating the mean of the distribution assume that

$$\mathrm{P}(X \geqslant 4)$$

can be approximated by

$$\mathrm{P}(X = 4).$$

Solution: We assume that $X \sim \text{Poisson}(\lambda)$, where λ needs to be determined from the data. We may estimate λ by computing the mean of the observations

$$\overline{X} = \frac{0 \cdot 20 + 1 \cdot 40 + 2 \cdot 16 + 3 \cdot 18 + 4 \cdot 6}{100} = 1.5.$$

With 1.5 as an estimate for λ, we can compute the predicted probabilities for each of the categories A_1, A_2, A_3, A_4, A_5 corresponding to $x = 0$, $x = 1$, $x = 2$, $x = 3$ and $x \geqslant 4$, respectively, using the pmf for a Poisson(λ) distribution:

$$p_X(x) = \frac{\lambda^x}{x!} \, e^{-\lambda} \quad \text{for} \ \ x = 0, 1, 2, \ldots,$$

and 0 otherwise. The predicted probabilities are shown in the second column of Table 4. Table 4 also shows the expected counts in each category in the third column.

A_i	p_i	Expected Counts	Observed Counts
$x = 0$	0.2231	22.31	20
$x = 1$	0.3347	33.47	40
$x = 2$	0.2510	25.10	16
$x = 3$	0.1255	12.55	18
$x \geqslant 4$	0.0657	6.57	6

Table 4: Data for Problem 5

Based on the data shown in the third and fourth column in Table 4, we compute the value of the Pearson Chi–Square statistic, Q, to be in this case

$$\widehat{Q} \approx 7.23$$

$\square$

(b) How many degrees of freedom are associated with the chi–square distribution used in this test?

Answer: Since we estimated one parameter and there are 5 categories, the numbers of degrees of freedom is df $= 5 - 1 - 1 = 3$.
$\square$

(c) Do the data support the rejection of the Poisson model at the $\alpha = 0.05$ significance level?

Solution: Since Q has an approximate $\chi^2(3)$ asymptotic distribution, we get the approximate p–value

$$p\text{–value} = \mathrm{P}(Q \geqslant \overline{Q}) \approx 0.0568,$$

which is bigger than $\alpha = 0.05$. Therefore, cannot reject the Poisson model at the 0.05 significance level. $\square$